Assessing Student Learning

Richard J. Shavelson is the Margaret Jacks Professor of Education, a professor of psychology, senior fellow in the Stanford Institute for the Environment, and former dean of the School of Education at Stanford University. A former president of the American Educational Research Association, he is completing a book tentatively titled The Quest to Assess Learning and Hold Higher Education Accountable. He gratefully acknowledges the support of the Atlantic Philanthropies for the work reported here.

Over the past 35 years, the public, as well as state and federal policy makers, have increasingly pressured higher education to create a culture of evidence and, specifically, to account for student learning. While virtually all states report on collegiate learning using proxies (e.g., graduation rates), colleges and universities are now being asked to assess learning directly. U.S. Secretary of Education Margaret Spellings' Commission on the Future of Higher Education, for example, has called for standardized tests of students' capacity to think critically, solve problems, and communicate analytically and clearly.
The commission has opened a conversation that we in higher education need to continue—a conversation about what our students should be learning and what the history of measurement tells us about how we should (and shouldn’t) determine whether they have learned. Here I would like to suggest what some of those goals might be, review that history, and end by suggesting how we might approach the task of assessing the full range of collegiate learning.

**LEARNING OUTCOMES**

There is more to be learned in college than the broad abilities singled out by the Secretary’s commission. The full range of outcomes for college students goes from factual, procedural, and conceptual knowledge and reasoning in a discipline, to thinking that might be applied to a very wide variety of situations, to the development of intelligence. Moreover, “cognitive” outcomes include “soft skills,” in that reasoning is involved in personal relations, moral challenges, and civic engagement. Chart 1 suggests the interrelationships among these kinds of learning.
If the learning outcomes of higher education are narrowly measured, as cost, capacity, and convenience would dictate, we risk narrowing the missions, subject matter taught, and diversity of the American system of higher education. For example, it is unsatisfactory to say that we should not measure soft skills because “the present state of the art in assessing these skills is not adequate for supporting the institution of a nationwide set of standardized measures” (Dwyer, Millett, & Payne, 2006, p. 20). If we do not assess them, these crucial skills will drop from sight as pressures for accountability lead campuses to focus on a more restricted, more easily and less expensively measured subset of learning outputs, such as multiple-choice tests of declarative content knowledge. What we need to do is to learn from the history of assessing learning and take audacious steps to develop and measure the full set of learning outcomes that our nation’s public so highly values.

This demand to establish a culture of evidence has a long lineage, and the future of this culture may depend on how well we understand the past. The key lesson we learn from that history is this: Cultures of evidence do not automatically lead to educational improvement if what counts as evidence does not also count as education or counts as only part of it.

**A BRIEF HISTORY OF LEARNING ASSESSMENT**

Spurred by the success of standardized “objective” mental testing in World War I (namely the Army Alpha test with multiple-choice and true-false questions having one correct answer), the first third of the 20th century marked the beginning of testing learning in higher education. During this period, the emphasis was on the mastery of academic content.

The Carnegie Foundation for the Advancement of Teaching led the movement. Howard Savage, a staff member and historian of the foundation in its early days, attributes Carnegie’s leadership in assessing college students’ learning to its first president, Henry Pritchett, who was motivated by his concern for the quality of higher education and his recognition of the potential impact that the emergence of “objective testing” might have on monitoring that quality. Walter A. Jessup, the foundation’s third president, later put what had become the foundation’s vision this way:

> [T]he central problems [in improving higher education] are three in number: first, the setting up of generally accepted standards of achievement; secondly, the devising of methods of measuring this achievement and holding pupils to performance; and thirdly, the introduction of such flexibility in educational offerings that each indi-
Pritchett’s passion was shared by his chief staff member, William S. Learned—“a man who had clear and certain opinions about what education ought to be ... [with] transmission of knowledge as the *sine qua non*” (Lagemann, 1983, p. 101). Among the ways Learned enacted that passion was a large-scale assessment in the state of Pennsylvania, in which 70 percent of college seniors in the state spent 12 hours taking a 3,200-item objective test, largely of declarative (i.e., facts and concepts) and procedural (i.e., “how-to”) content knowledge.

Objective testing followed directly from the study’s conception of collegiate learning as the accumulation of content knowledge, since it could more efficiently and reliably verify the accumulation of that knowledge than could the then-widely used essay examination. E. L. Thorndike’s study of engineering students during this period made a similar assumption. Thorndike administered objective tests to students at the Massachusetts Institute of Technology, the University of Cincinnati, and Columbia University concerning facts, concepts, and mathematical routines.

The Pennsylvania study proved that comprehensive assessment of content knowledge was feasible. The most noteworthy next step taken in testing during the progressive period between 1933 and 1947 was the attempt to assess not only cognitive outcomes beyond factual and procedural knowledge, but also the personal, social, and moral outcomes of general education.

During this period, general education and general colleges, such as the University of Chicago’s progressive College, sprouted in universities across the country. In the Chicago program, a central University Examiner’s Office, rather than the faculty, was responsible for developing, administering, and scoring tests in the university’s general-education program. The Chicago examinations tested a much broader range of knowledge and abilities than the Pennsylvania study had: the use of knowledge in a variety of unfamiliar situations; the ability to apply principles to explain phenomena; and the ability to predict outcomes, determine courses of action, and interpret works of art. Open-ended essays and multiple-choice questions demanding interpretation, synthesis, and application of new texts (primary sources) characterized the comprehensive exams. The ubiquitous “Bloom’s Taxonomy” of cognitive objectives (knowledge, comprehension, application, analysis, synthesis, evaluation) grew out of the work of the Examiner’s Office.

Meanwhile, the Cooperative Study of General Education, conducted by a consortium of higher-education institutions, reflected progressive notions of human development as well. Members of the consortium believed they could benefit from a cooperative attempt to improve general education. To this end they assessed students’ achievement and well-being. The consortium developed the Inventory of General Goals in Life, the Inventory of Satisfactions Found in Reading Fiction, the Inventory of Social Understanding, and the Health Inventories.

But even as these expanded notions of learning were being acted on, Learned parlayed his experience with the Pennsylvania study into an assessment for graduate education. In proposing the “Cooperative Graduate Testing Program,” Learned noted that with increased demand for graduate education following the Depression and with the decreasing credibility of the baccalaureate degree, something more than the number of college credits a student had earned was needed on which to base graduate admissions decisions.

In concert with the graduate schools at Columbia, Harvard, Princeton, and Yale, in 1937 Learned’s team administered seven tests—the first administration of what was to be the Graduate Record Examination (GRE). From there the program grew by leaps and bounds (Chart 2, p. 30). Despite its success, the growing financial and logistical burdens it imposed led the Carnegie Foundation to spin off an independent national testing agency, the Education Testing Service (ETS), to administer the exam.

The original GRE, like the Pennsylvania study’s examination, was a comprehensive, objective test focused largely on students’ content knowledge, but it also tapped verbal reasoning and was used to infer students’ fitness for graduate study. In 1949 the GRE also began to test general reasoning via the GRE Aptitude Test, with the verbal and quantitative sections we see today.

ETS continued the shift away from content toward general reasoning in 1954 by replacing the profile tests and
the tests of general education with “area
tests” that focused on the social and
atural sciences and humanities, as a
means of assessing the broad outcomes
of the liberal arts. The tests emphasized
reading comprehension, understanding,
and interpretation, often providing the
requisite content knowledge “because of
the differences among institutions with
regard to curriculum and the differences
among students with regard to specific
course selection” (ETS, 1966, p. 3).

When the Carnegie Foundation
moved the GRE to ETS, it left an ex-
traordinarily strong legacy that has
endured into the 21st century: objective,
group-administered, cost-efficient test-
ing using selected response—now solely
multiple-choice—questions. Testing
organizations developed precursors to
today’s major learning-assessment pro-
grams in the 1960s and 1970s, including
ETS’s Undergraduate Assessment
Program, which incorporated the GRE,
and American College Testing’s College
Outcomes Measures Project (COMP).
The former evolved via the Academic
Profile into today’s Measure of Acad-
emic Proficiency and Progress (MAPP)
and the latter into today’s Collegiate
Assessment of Academic Proficiency
(CAAP)—multiple-choice test batteries
designed to measure and improve gen-
eral-education outcomes.

However, several legacies of the pro-
gressive era emerged in the late 1970s,
auguring a change in the course set by
Learned. Faculty members were not en-
tirely happy with multiple-choice tests.
They wanted to get at broader abilities,
such as the ability to communicate,
think analytically, and solve problems.
This led to several developments. ETS
studied “constructed-response” tests
that tapped communication skills, ana-
lytic thinking, synthesizing ability, and
social/cultural awareness, while Ameri-
can College Testing (ACT) tried out
open-ended, performance-based assess-
ments of skills for effective functioning
in adult life. And New Jersey developed
Tasks in Critical Thinking, which asked
students to perform real-world tasks in
a “performance-based assessment ...
[off] the skills of inquiry, analysis, and
communication” (ETS, 1994, p. 2), with
prompts that did “not assess content or
recall knowledge” (p. 2).

For a short period of time, these as-
sessment programs set the mold. But
due to time limitations, as well as is-
ues with scoring, reliability, and cost,
they either faded or morphed back into
multiple-choice tests. By the beginning
of the 1980s, political pressure to assess
student learning and hold campuses ac-
countable for it led to an increased de-
mand for standardized, multiple-choice
testing (e.g., in Florida and Tennessee).

By the mid-80s, though, some states
(e.g., Virginia, South Carolina, and
Missouri) had mandated not standard-
ized testing but campus-based assess-
ment, to which the campuses responded
by creating individualized assessment
programs. Some of those programs re-
lied on home-grown assessments, while
others looked to the testing companies to
provide measures. By this time, a wide
array of college-learning assessments in
the Carnegie tradition was available,
which remain popular to this day.

In the area of general education, ETS
currently provides the MAPP and ACT
administers the CAAP. The College
Resource Center at the University of
Missouri, Columbia, offers the College
Basic Academic Subjects Examination
(CBASE), a criterion-referenced achieve-
ment examination in English, mathemat-
ics, science, and social studies that
serves to qualify individuals for entry into teacher-education programs and as a test of
general academic knowledge and skills.

But as we saw at the end of the 1970s,
objective testing was not the way faculty
members wanted student learning to be
assessed. They were more comfortable
with open-ended, holistic, problem-
based assessments, which were more in
tune with what they thought they were
teaching. Intuitively, they suspected that
the kind of thinking that is stimulated
and the type of performance that is as-
essed by multiple-choice and other
highly structured tests differ from that
generated by more open-ended tasks.

And empirical evidence supports
their intuition. While a multiple-choice
test and a “constructed-response” test
may produce scores that are correlated
with each other, this correlation does not
mean that the same kind of reasoning
is involved. Students’ performance var-
ies considerably when the same task is
presented as a multiple-choice question,
an open-ended question, or a concrete
performance task.

The current (November 2006) draft
report of the Secretary of Education’s
higher-education commission specific-
ally mentioned two tests as examples
when it recommended that “postsec-
ondary education institutions should
measure and report meaningful student
learning outcomes”: the MAPP and the
College Learning Assessment (CLA).
In supporting the commission’s recom-
mandation, the American Association
of State Colleges and Universities zero-
ed in on the CLA, giving a rationale
for that choice that echoes the progres-
sive-era arguments that are consonant
with faculty preferences and with best
practices in teaching and learning.

**THE COLLEGIATE LEARNING ASSESSMENT**

The Council for Aid to Education
developed the Collegiate Learning.

**CHART 2. GRE’S GROWTH OVER ITS FIRST DECADE**

![Chart 2](chart-image-url)
Assessment (CLA). As the CLA may provide insight into a next generation of learning assessments, I note its major features here (see table, p. 32).

Just as objective testing and computer scoring of response sheets (developed by IBM for Learned) revolutionized assessment of learning at the turn of the 20th century, so too has new information technology and statistical sampling ushered in a significant change in assessing college students’ learning at the turn of the 21st century. The CLA moves away from a selected-response, multiple-choice strategy to one that test developers from the progressive era would have recognized—asking students to perform concrete, complex, open-ended tasks.

The CLA’s notions of learning, as well as of assessment, had their origin in that same period. The CLA assesses not content knowledge but the skills of critical thinking, analytic reasoning, problem solving, and written communication. These capabilities are tapped in realistic “work-sample” tasks drawn from education, work, and everyday life. They are accessible to students from a wide diversity of majors and general-education programs.

The assessment is divided into two parts (plus biographical information): analytic writing and performance tasks (for examples of both, go to http://www.cae.org/content/pdf/CLA.in.Context.pdf). Two types of writing tasks are administered. The first, “make an argument,” invites students to present an argument for or against a particular position. The second, “break an argument,” asks students to critically evaluate an argument.

The performance tasks present real-life problems to students, along with resources to solve them, such as an “in-basket” (or nowadays, “computer basket”) with packets of information, some of which bear on the problem and some of which are irrelevant. (Part of the assessment is for the students to decide what information to use and what to ignore.) Students integrate these multiple sources of information to arrive at a solution, decision, or recommendation. Students respond in realistic ways—for example, by writing a memorandum analyzing the pros and cons of alternative solutions, and recommending what a hypothetical company should do. In scoring performance, alternative justifiable solutions to the problem and alternative paths to solving it are recognized and evaluated.

The test can be scored reliably without incurring the prohibitive costs of traditional essay testing due to recent developments in information technology. The assessment is delivered on an interactive Internet platform that produces a paperless, electronic administration and online reports of results. Written communication tasks are scored using natural-language processing software. Performance tasks, currently scored by human raters, will also be scored by computer software within a year.

Sampling is another key feature that makes CLA-type tests feasible. Students are no longer willing to sit for 12-hour exams, as they once did in Pennsylvania. The capacity to provide these rich tasks without overburdening students or increasing costs to unsustainable levels is a function of statistical sampling (all students need not be tested on all tasks), which is made possible by the different goals of the testing. The basic focus in Pennsylvania was on individual students’ development; CLA focuses on program improvement.

Institutional (and subdivision) reports provide a number of indicators for interpreting performance, including anonymous benchmark institutional comparisons and the percentage of institutions scoring below a certain level. The CLA also enables institutions to determine the value they add to students’ learning, over and above the performance expected at the institution based on the abilities of admitted students, cross-sectional comparisons, longitudinal cohort studies, or some combination.

Chart 3 (p. 33) shows the performance of a set of colleges’ freshmen and seniors in spring 2005. Each point on the graph represents average (mean) college performance on the ACT/SAT and the CLA; the green dots and line are seniors, and the purple dots and line are freshmen.

A number of features in Chart 3 are noteworthy. First, perhaps most encouragingly, the green dots and line (seniors) fall significantly (more than 1.6 standard deviations) above the purple dots and line (freshmen), which suggests that college does indeed contribute to student learning (as do other life experiences). Second, most dots fall along the straight (“regression”) line of expected performance based on ability for both freshmen and seniors—but
some fall well above and some well below. This means that by the senior year some colleges exceed expected performance compared to their peers, and some perform below expectation. So it matters not only that you go, but also where you go to college.

The CLA does not pretend to be the measure of collegiate learning. Rather, as the Council for Aid to Education points out, the CLA focuses on the broad abilities of critical reasoning, problem solving, and communication, which do not exhaust the list of collegiate learning outcomes. Moreover, with its institutional (or school/college) focus, it does not provide detailed, diagnostic information about particular courses or programs (unless sampling is done at a program level). Other institutional information is needed to diagnose problems, and campuses need to test possible solutions to those problems systematically. The CLA, then, sends a strong signal to the campus to dig deeper.

AN AUDACIOUS PROPOSAL FOR ASSESSING LEARNING RESPONSIBLY

Why not use a CLA-type framework—with its focus on broad cognitive abilities embedded in meaningful, holistic, complex tasks and its use of information technologies—to assess academic programs? That is, why not shape a writing task around, for example, an historical event? History students could be given a computer basket of information on the event and asked to adjudicate among competing interpretations of what happened. It might be possible for disciplinary societies, in this case the American Historical Association, to work with the Council for Aid to Education to create measures that tap knowledge and reasoning within the discipline.

To be sure, this would be a challenging task. Issues of how much factual and conceptual historical knowledge, and in what domains of history to require for successful completion of the task would be hotly debated. But the writing and performance tasks would not be designed to elicit students’ declarative content knowledge so much as to embody the kinds of thinking that faculty want their students to develop within the major. For that purpose, both real and fictitious events, or a combination, could be used. Having learned societies engage in building assessments with an assessment-development organization would create learning for both partners, create some disciplinary buy-in, and build the capacity of the society to assist professors in designing similar tasks for their courses.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open-ended Tasks</td>
<td>• Critical thinking, analytic reasoning, problem-solving and written communication are assessed</td>
</tr>
<tr>
<td></td>
<td>• Realistic work samples</td>
</tr>
<tr>
<td></td>
<td>• Engaging tasks</td>
</tr>
<tr>
<td></td>
<td>• Applicable to different academic majors</td>
</tr>
<tr>
<td>Computer Technology</td>
<td>• Interactive Internet platform</td>
</tr>
<tr>
<td></td>
<td>• Paperless administration</td>
</tr>
<tr>
<td></td>
<td>• Natural-language processing software for scoring writing</td>
</tr>
<tr>
<td></td>
<td>• Human scoring of performance tasks, currently</td>
</tr>
<tr>
<td></td>
<td>• Report of institution's (and subdivision's) performance (and individual student performance confidentially to student)</td>
</tr>
<tr>
<td>Focus</td>
<td>• Institutional or school/department/program</td>
</tr>
<tr>
<td></td>
<td>• Not on individual student performance</td>
</tr>
<tr>
<td>Sampling</td>
<td>• Not all students perform all tasks</td>
</tr>
<tr>
<td></td>
<td>• Creates scores at institution or subdivision/program level as desired (depending on sample sizes)</td>
</tr>
<tr>
<td>Reporting</td>
<td>• Controls for student ability so that &quot;similarly situated&quot; benchmark campuses can be compared</td>
</tr>
<tr>
<td></td>
<td>• Provides value-added estimates</td>
</tr>
<tr>
<td></td>
<td>• Provides percentiles</td>
</tr>
<tr>
<td></td>
<td>• Provides benchmark institutions</td>
</tr>
</tbody>
</table>

Source: Richard J. Shavelson
CLA-type measures could also be used to assess what the AAC&U calls individual and social-responsibility outcomes. Indeed, some of the tasks on the current CLA tap some of these competencies. Why not create more such tasks—for instance, one that focuses on a civic issue, perhaps regarding the right to freedom of speech at a time when a fictitious country is engaged in a contentious war or about the effectiveness of a group of local environmentalists’ argument for removing an old dam?

CONCLUSION

Today's demand to establish a culture of evidence about student learning appears to be new, but it is not. What is clear from this review of past and current attempts to assess college students' learning is that cultures of evidence do not lead to educational improvement if what counts as evidence does not count as education or only counts as a part of what we expect from a college education.

Over the past 100 years, two notions about what college students should learn and how they should be tested have competed: the behaviorist notions of achievement reflected in Carnegie's leadership and embedded in the MAPP, CAAP, and CBASE have vied with the progressive notions of learning that we see in the University of Chicago's general-education assessment, the Cooperative Study's personal and social-development tests, and attempts to assess general intellectual skills such as the Tasks in Critical Thinking and now the CLA. No doubt these views will continue to tussle for the next century. That said, my bet is on CLA-type assessments.

The history of assessing learning provides some important lessons. It is important that we develop and justify a conceptual framework for college outcomes and assessment of achievement and learning. We must design assessment systems to collect both snapshots of performance at one point in time (achievement) and over time (learning). We need to include in those assessments the broad spectrum of collegiate outcomes, including the difficult-to-measure personal and social perspectives that we value so highly, since what we measure is what we will come to teach.

CHANGE • JANUARY/FEBRUARY 2007

CHART 3. RELATIONSHIP BETWEEN MEAN SAT (AND ACT CONVERTED) AND CLA SCORES

Finally, we need to recognize that any large-scale assessment system can, at best, signal where a problem may exist, but it will not pinpoint the problem and generate solutions. For this, a campus needs to place any external assessment in the context of its own rich array of assessments and link it to institutional structures and processes. Only by doing this will it be able to bring assessment information to bear on the improvement of teaching and learning.

RESOURCES